CLAIMS

What is claimed is:

1. A method for managing admission of requests to a shared media server, the method comprising:

allowing each of a plurality of hosting services access to any of a set of shared resources for serving their respective streaming files to clients; and

managing admission of client requests for streaming files to each of the plurality of hosting services to ensure that a desired amount of usage of the shared resources is available to each hosting service.

2. The method of claim 1 further comprising:

implementing the plurality of hosting services on a shared media server.

3. The method of claim 1 wherein the set of shared resources comprises:

shared memory resources and shared disk resources.

4. The method of claim 1 further comprising:

determining the desired amount of usage of the shared resources for a hosting service from a service level agreement.

5. The method of claim 1 wherein said set of shared resources comprises memory and wherein said managing admission of client requests for streaming files comprises:

receiving a client request for a streaming file to be served from one of said hosting services; and

using a segment-based memory model to determine whether at least a portion of the requested streaming file is in the memory.

6. The method of claim 5 further comprising:

determining from the segment-based memory model a cost associated with the one of said hosting services serving the requested streaming file.

7. The method of claim 1 wherein said managing admission of client requests for streaming files comprises:

receiving a new request for service of a streaming file by one of the plurality of hosting services;

performing a resource availability check for the one of a plurality of hosting services to determine whether the requested hosting service has sufficient available resource usage allocated thereto to service the new request.

8. The method of claim 7 wherein said managing admission of client requests for streaming files further comprises:

performing a performance isolation guarantee check for the plurality of hosting services to determine whether acceptance of the new request will violate, at any point in the future, availability of a desired amount of usage of the shared resources for any of the plurality of hosting services.

9. The method of claim 1 further comprising:

specifying, for each of the plurality of hosting services, a desired amount of usage of the shared resources to be available at any given time for the hosting service.

10. The method of claim 9 wherein said managing admission of client requests comprises:

managing admission of client requests for streaming files to each of the plurality of hosting services to ensure that each of the plurality of hosting services has usage of its corresponding specified desired amount of the shared resources.

11. A system comprising:

a media server comprising a plurality of hosting services for streaming files implemented thereon, wherein the media server comprises shared resources and wherein the plurality of hosting services share usage of the media server's shared resources in serving streaming files to their respective clients; and

an admission controller for managing admission of client requests for service to each of the plurality of hosting services to ensure that no one of the plurality of hosting services overtakes usage of an undesirably high proportion of the shared resources.

12. The system of claim 11 wherein the admission controller is operable to manage admission of said client requests to the plurality of hosting services to ensure that a desired amount of usage of the shared resources is available, at any given time, to each hosting service.

13. The system of claim 11 wherein the shared resources comprise memory resources and disk resources.

14. The system of claim 13 wherein the admission controller is operable to use a segment-based model of the memory resources to determine whether at least a portion of a requested streaming file is in the memory resources.

15. The system of claim 11 wherein said admission controller is operable to receive a new request for service of a streaming file by one of the plurality of hosting services, and determine whether the requested hosting service has sufficient available resource usage allocated thereto to service the new request.

16. The system of claim 15 wherein an amount of resource usage is preallocated to the requested hosting service.

17. The system of claim 15 wherein said admission controller is further operable to determine whether acceptance of the new request will violate, at any point in the future, availability of a desired amount of usage of the shared resources for any of the plurality of hosting services.

18. A method for managing admission of requests to hosting services that share resources, the method comprising:

allowing each of a plurality of hosting services access to any of a set of shared resources for serving their respective files to clients thereof;

for each of the plurality of hosting services, identifying a desired amount of usage of the set of shared resources to be available for the hosting service; and

isolating usage of the set of shared resources by the plurality of hosting services to ensure that the respective desired amount of usage of the set of shared resources is available to each hosting service.

19. The method of claim 18 wherein said hosting services host streaming files for access by clients thereof.

20. The method of claim 18 wherein the set of shared resources comprise:

shared memory resources and shared disk resources.

21. The method of claim 18 further comprising:

determining the desired amount of usage of the set of shared resources for a hosting service from a service level agreement.

22. The method of claim 18 wherein said set of shared resources comprises memory and wherein said isolating usage of the set of shared resources comprises:

specifying, for each of the hosting services, an amount of usage of the set of shared resources to be available, at any time, to the hosting service; and

determining whether acceptance of a new request for service by a hosting service will violate, at any point in the future, availability of a specified amount of usage of the shared resources for any of the plurality of hosting services.

23. A method for managing admission of requests to a hosting service, the method comprising:

allowing each of a plurality of hosting services access to any of a set of shared resources for serving their respective files to clients thereof;

for each of the hosting services, identifying a desired amount of usage of the set of shared resources to be available for the hosting service;

receiving a new request for a streaming file to be served by one of the hosting services;

determining a cost to the one of the hosting services for serving the requested streaming file, wherein the cost corresponds to the shared resources to be consumed in serving the requested streaming file; and

determining, based at least in part on the cost, whether to admit the new request for service by the one of the hosting services.

24. The method of claim 23 wherein said determining whether to admit the new request comprises:

determining whether the cost exceeds the amount of usage of the shared resources allowed for the one of the hosting services.

25. The method of claim 23 wherein said determining whether to admit the new request comprises:

determining whether the cost of shared resources consumed violates availability of a desired amount of usage of the set of shared resources to be available for another one of the hosting services.

26. The method of claim 23 wherein the set of shared resources comprises:

shared memory resources and shared disk resources.

27. The method of claim 23 wherein said determining whether to admit the new request comprises:

determining whether the requested hosting service has sufficient available resource usage allocated thereto to service the new request; and

determining whether acceptance of the new request will violate, at any point in the future, availability of a desired amount of usage of the shared resources for any of the plurality of hosting services.

28. A method comprising:

allowing each of a plurality of hosting services access to any of a set of shared resources for serving their respective files to clients thereof, wherein the shared resources includes a memory;

receiving, at a time T_{cur}, a new request for a streaming file to be served by one of the hosting services;

creating a segment-based model of the memory as of time T_{cur}; and

based at least in part on the segment-based model of the memory, determining whether to accept the received request for service by the hosting service.

29. The method of claim 28 further comprising:

for each of the plurality of hosting services, identifying a desired amount of usage of the set of shared resources to be available at any time for the hosting service.

30. The method of claim 28 wherein said determining whether to accept the received request for service by the hosting service comprises:

determining whether the requested hosting service has sufficient available resource usage allocated thereto to service the received request; and

determining whether acceptance of the received request for service by the requested hosting service will violate, at any point in the future, availability of a desired amount of usage of the shared resources for any of the plurality of hosting services.

31. The method of claim 28 wherein said segment-based model of the memory comprises (a) identification of unique segments of streaming files previously accessed by clients and (b) identification of corresponding timestamps of most recent accesses of each unique segment.

32. Computer-executable software stored to a computer-readable medium, the computer-executable software comprising:

code for creating a segment-based model of a media server's memory, wherein the media server's memory is a shared resource to which a plurality of hosting services implemented on the media server have access for serving their respective files to clients thereof; and

code for determining whether to serve a requested streaming file from one of the plurality of hosting services based at least in part on the segment-based model of the media server's memory.

33. The computer-executable software code of claim 32 wherein said code for determining whether to serve a requested streaming file from one of the plurality of hosting services comprises:

code for determining whether the one of the plurality of hosting services has sufficient available resource usage allocated thereto to serve the requested streaming file; and

code for determining whether acceptance of the received request for service by the one of the plurality of hosting services will violate, at any point in the future, availability of a desired amount of usage of the shared resources for any of the plurality of hosting services.

34. The computer-executable software code of claim 32 wherein said segment-based model of the media server's memory comprises (a) identification of unique segments of streaming files previously accessed by clients of the media server and (b) identification of corresponding timestamps of most recent accesses of each unique segment.

35. The computer-executable software code of claim 32 wherein said code for determining whether to serve a requested streaming file from one of the plurality of hosting services comprises:

code for determining a cost to the one of the plurality of hosting services for serving the requested streaming file, wherein the cost corresponds to the amount of the shared resources to be consumed in serving the requested streaming file.

36. An admission controller for managing admission of requests to hosting services that share resources, the admission controller comprising:

means for receiving a new request for a streaming file to be served by one of a plurality of hosting services that share access to a set of shared resources for serving their respective files to clients thereof;

means for performing a resource availability check for the one of a plurality of hosting services from which the streaming file is requested by the new request to determine whether the requested hosting service has sufficient available resource usage allocated thereto to service the new request; and

means for performing performance isolation guarantee check for the plurality of hosting services to determine whether acceptance of the new request will violate, at any point in the future, availability of a desired amount of usage of the shared resources for any of the plurality of hosting services.

37. The admission controller of claim 36 wherein said means for performing a resource availability check comprises:

means for determining a cost associated with the one of a plurality of hosting services serving the requested streaming media file, wherein the cost corresponds to the shared resources to be consumed in serving the requested streaming file.

38. The admission controller of claim 36 wherein said set of shared resources comprises:

memory and disk resources.